BIOLOGICAL CRITERIA
Technical Guidance for Survey Design and Statistical Evaluation of Biosurvey Data

CHAPTER 4 Detecting Mean Differences

Hypothesis testing methods that seek to detect the mean differences arising from two or more independent samples are among the most common statistical procedures performed. However, these procedures are frequently used without regard to some basic assumptions about the data under investigation — which, in some cases, leads to errors in interpretation.

This section describes and illustrates several methods for detecting mean differences. It focuses on (1) cases in which only two means are involved, and (2) situations involving more than two means. It also presents suggestions concerning the use and abuse of means testing procedures.

Cases Involving Two Means

Several scenarios within the biocriteria program require investigators to compare the mean differences between two independent populations. Suppose for example, that we want to use biocriteria in a regulatory setting in the following situation:

A wastewater treatment plant discharges its effluent into a stream at a single point. Upstream of the discharge facility, the stream is in good shape (unaffected by any known sources of pollution). The resource agency has sufficient funds to monitor three stations upstream of the discharge site and a comparable number of streams downstream of the discharge site during the late summer. The agency has chosen to evaluate aquatic life use impairment using benthic species richness.

At each of the six sites, 10 independent measures of species richness were generated by randomly placed ponar grabs over a relatively small spatial area (a sample size of 10 was chosen based on variability estimates generated at a different, but similar site). Sites of comparable habitat quality were chosen for sampling. The upstream sites will serve as a reference condition against which to compare the downstream condition.

In addition to the current survey (i.e., sampling regime, data collection, and interpretation), the regulatory agency has identified an additional upstream site for which it has 10 years of comparable long-term (historical) data. The investigators have no reason to believe that a time component exists in the long-term data. Table 4.1 presents descriptive information associated with the upstream and downstream sites and with the long-term site.

The question for investigators is this: Do the data reveal a downstream effect associated with the wastewater discharge? Several methods are available for assessing the mean differences between the upstream and downstream sites, and each method has both positive and negative aspects.

Random Sampling Model, External Value for σ

Suppose investigators believe that the 30 measures of benthic species richness collected at the upstream and downstream sites can be treated as random samples from appropriate populations. In particular, they

Table 4.1—Descriptive statistics: upstream-downstream measures of benthic species richness.

SITE		*N*	MEAN	STD.	MINIMUM	MAXIMUM	10%–TRIMMED MEAN	MEDIAN ABSOLUTE DEVIATION
Upstream	1	10	10.0	2.3	7.5	14.8	9.7	1.5
	2	10	12.6	2.5	10.3	18.0	12.2	1.3
	3	10	11.2	2.4	7.2	15.1	11.2	1.0
Downstream	4	10	10.4	2.4	6.3	13.7	10.5	1.0
	5	10	7.7	3.7	3.4	14.7	7.4	2.7
	6	10	9.0	1.8	5.6	11.1	9.1	1.5
Historic	7	200	10.4	3.4	0.17	19.4	11.1	2.6
Pooled Data	1–3	30	11.3	2.5	7.2	18.0	11.1	1.0
	4–6	30	9.0	2.9	3.4	14.7	9.0	1.6

believe that the two populations have the same form (i.e., normal distributions with the same variance, σ) but different means, μ_a and μ_b. How can the investigators use statistical theory to make inferences about the effect of the wastewater treatment plant discharge?

If the data were random samples from the populations, with N_a = 30 observations from the upstream population and N_b observations from the downstream population, the variances of the calculated averages, Y_a and Y_b would be:

$$V(\mathrm{Y_a}) = \frac{\sigma^2}{\mathrm{N_a}} \quad , \quad V(\mathrm{Y_b}) = \frac{\sigma^2}{\mathrm{N_b}} \tag{4.1}$$

Likewise, in the random sampling model, Y_a and Y_b would be distributed independently, so that:

$$V(\mathrm{Y_a} - \mathrm{Y_b}) = \frac{\sigma^2}{\mathrm{N_a}} + \frac{\sigma^2}{\mathrm{N_b}} = \sigma^2\left(\frac{1}{\mathrm{N_a}} + \frac{1}{\mathrm{N_b}}\right) \tag{4.2}$$

Even if the distributions of the original observations had been moderately nonnormal, the distribution of the difference Y_a–Y_b between sample averages would be nearly normal because of the central limit effect. Therefore, on the assumption of random sampling,

$$z = \frac{(\mathrm{Y_b} - \mathrm{Y_a}) - (\mu_b - \mu_a)}{\sigma\sqrt{\frac{1}{\mathrm{N_a}} + \frac{1}{\mathrm{N_b}}}} \tag{4.3}$$

would be approximately a unit normal deviate.

Now, σ, the hypothetical population value for the standard deviation, is unknown. However, the historical data yield a standard deviation of 3.4. If this value is used for the common standard deviation of the sampled populations, the standard error of the difference, Y_a-Y_b = 2.3, is

$$\sigma\sqrt{\frac{1}{30} + \frac{1}{30}} = 0.89$$

Based on the robust estimators (trimmed mean difference of 2.1 and median absolute difference of 1.6) the standard error of the difference would be 0.41. If the assumptions are appropriate, the approximate significance level associated with the postulated difference (μ_a–μ_b) in the population means will then be obtained by referring

$$z_0 = \frac{2.3 - (\mu_a - \mu_b)_0}{.89}$$

to a table of significance levels of the normal distribution. In particular, for the null hypothesis (μ_a–μ_b) = 0, z_0 = 2.3/.89 = 2.6, and Pr(z < 2.6) < .005. Again, the upstream/downstream effect seems to be realistic (using the robust estimators, z = 5.1 and Pr[z < 5.1] < .001). Note that we use the *z* distribution in this example because the population variance is determined from an external set of data that represents the population of interest — an assumption equivalent to assuming that the variance of the population is known (i.e., not estimated).

Random Sampling Model, Internal Value for σ

Suppose now that the only evidence about σ is from the N_a = 30 samples taken upstream and the N_b = 30 samples taken downstream. The sample variances are

$$s_a^2 = \frac{\sum(\mathrm{Y_{a1}} - \mathrm{Y_a})^2}{\mathrm{N_a} - 1} = 6.25$$

$$s_b^2 = \frac{\sum(\mathrm{Y_{b1}} - \mathrm{Y_b})^2}{\mathrm{N_b} - 1} = 8.41$$

On the assumption that the population variances of the upstream and downstream sites are, to an adequate approximation, equal, these estimates may be combined to provide a pooled estimate of s^2 of this common σ^2. This is accomplished by adding the sums of squares in the numerators and dividing by the sum of the degrees of freedom,

$$s^2 = \frac{\sum(\mathrm{Y_{a1}} - \mathrm{Y_a})^2 + \sum(\mathrm{Y_{b1}} - \mathrm{Y_b})^2}{\mathrm{N_a} + \mathrm{N_b} - 2} = 7.52$$

On the assumption of random sampling from normal populations with equal variances, in which the discrepancy [(Y_a–Y_b) – (μ_a–μ_b)] is compared with the estimated standard error of Y_a–Y_b, a *t* distribution with N_a+N_b–2 degrees of freedom is appropriate. The *t* statistic in this example is calculated as

$$t = \frac{(\mathrm{Y_a} - \mathrm{Y_b}) - (\mu_a - \mu_b)}{s\sqrt{\frac{1}{\mathrm{N_a}} + \frac{1}{\mathrm{N_b}}}} = \frac{2.31}{0.71} = 3.2$$

This statistic is referred to a *t* table with 58 degrees of freedom. In particular, for the null hypothesis that (μ_a–μ_b) = 0, Pr(*t* < 3.2) < .001. Again, an upstream/downstream effect seems plausible. Using the robust statistics, a pooled estimate of error can be calculated as the average of the median absolute deviations associated with each data set ([1 + 1.6] / 2 = 1.3). Therefore, the *t* statistic is 6.3 and the Pr(*t* < 6.3)< .001. Note that we use the *t* distribution in this example because the population variance is estimated from the survey data and not assumed to be known.

Testing against a Numeric Criterion

In the preceding sections, hypothesis tests were presented for the two-sample case. Similar tests are available for testing a sample mean against a fixed numeric criterion (for which an associated uncertainty does not exist). In this case, the *t* statistic can be written as follows:

$$t = \frac{Y - \mu}{s\sqrt{\frac{1}{n}}} \tag{4.4}$$

Here, *s* is the sample standard deviation and μ is the numeric criterion of interest. The probability of a greater value can be found in a *t* table using *n-1* degrees of freedom.

A Distribution-Free Test

In many instances, the assumption that the raw data (or paired differences) are normally distributed does not hold. Even the simplest monitoring design involving the comparison of two means requires either (1) a long sequence of relevant previous records that may not be available or (2) a random sampling assumption that may not be tenable. One solution to this dilemma is the use of distribution free statistics such as the *W* rank sum test (Hollander and Wolfe, 1973). The *W* test is designed to test the hypothesis that two random samples are drawn from identical continuous distributions with the same center. An alternative hypothesis is that one distribution is offset from the other, but otherwise identical. Comparative studies of the *t* and *W* tests indicate that while the *t* test is somewhat robust to the normality assumption, the *W* test is relatively powerful while not requiring normality. In many cases, performing both the *t* and *W* tests can be used as a double check on the hypothesis.

To conduct the *W* test (see Chapter 2), the investigator combines the data points from the samples, but maintains the separate sample identity. This overall data set is ordered from low value to high value, and ranks are assigned according to this ordering. To test the null hypothesis of no difference between the two distributions *f*(*x*) and *g*(*x*) (i.e., H_0: *f*[*x*] = *g*[*x*]), the ranks of the data points in one of the two samples are summed:

$$W = \sum R_i \tag{4.5}$$

Statistical significance is a function of the degree to which, under the null hypothesis, the ranks occupied by either data set differ from the ranks expected as a result of random variation. For small samples, the *W* statistic calculated in Equation 4.5 can be compared to tabulated values to determine its significance. Alternatively, for moderate to large samples, *W* is approximately normal with mean *E*(*W*) and variance *V*(*W*):

$$E(W) = \frac{N_a(N_b + N_a + 1)}{2} \tag{4.6}$$

$$V(W) = \frac{N_a N_b (N_b + N_a + 1)}{12} \tag{4.7}$$

$$z = \frac{W - E(W)}{\left[V(W)\right]^{0.5}} \tag{4.8}$$

In the upstream/downstream case that we have been discussing, *E*(*W*) = 1,127, *z* = 3.12, and Pr(<*z*)= 0.0018.

Evaluating Two-sample Means Testing

Table 4.2 summarizes the advantages and disadvantages of these two-sample means testing procedures. Both of these methods, to one degree of another, involve assumptions of normality, equality of variance, and independence. In all cases, the latter assumption is of greatest concern. Therefore, data with inherent time trends, seasonal cycles, or spatial correlations unrelated to the effect of interest should be carefully scrutinized prior to hypothesis testing using these procedures. Investigators can remove time trends and spatial correlations from the data prior to testing them for mean differences (Reckhow, 1983).

Multiple Sample Case

Hypothesis testing of multiple sample mean differences can be accomplished using both parametric (assumes normality) and nonparametric (no assumption of normality) approaches. The typical parametric approach to multiple means testing falls under the broad class of statistical models and methods called analysis of variance (ANOVA). Nonparametric counterparts include a number of specific tests including, among others, the Kruskal-Wallis rank sum test.

Both the parametric and nonparametric methods can be used with experimental and survey type data. However, the development of these statistical models include many permutations and assumptions and cannot be covered in this text. Instead, a brief discussion of each method is followed by an example of their typical outputs.

Parametric or Analysis of Variance Methods

ANOVA methods are a class of techniques for analyzing experimental data. A continuous response variable, known as the dependent variable, is measured under experimental conditions identified by classification variables known as independent variables or treatments. The variation in response is explained as an effect of the classification variable and random error.

Numerous decisions must be made by the investigator before attempting to use ANOVA procedures.

Table 4.2.—Assumptions, advantages, and disadvantages associated with various two-sample means testing procedures.

REFERENCE DISTRIBUTION	ASSUMPTIONS	ADVANTAGES	DISADVANTAGES	SHOULD CONSIDER FOR USE WHEN:	SHOULD NOT USE WHEN:
External	Past data can provide relevant reference set for observed difference Y_a–Y_b	No assumption of independence of errors. No need for random sampling hypothesis.	Need relevant, lengthy past records. Construction of reference distribution can be tedious	Quality, consistency, and length of data are deemed to represent a healthy ecosystem.	Known impacts to reference site have occurred, or physical and biological differences between the impact and reference site are identified.
Normal distribution with external estimate of σ	Individual observations are as if obtained by random sampling from normal populations with common standard deviation.	Continuous reference distribution that is easy to calculate.	Need to know σ. Need assumption of independence of individual errors coming from random sampling hypothesis.	Quality, consistency, and length of data are deemed to be a sample from a healthy ecosystem. Data transformation may be necessary to achieve normality.	Quality of data is suspect or impacts at the external site are known or suspected.
Normal distribution with internal estimate of σ	Individual observations are as if obtained by random sampling from normal populations with unknown common standard deviation σ estimated by s	No external data needed.	Need assumption of independence of individual errors coming from random sampling hypothesis.	Most commonly used test. Appropriate if normality assumptions hold. If outliers or influential data apprent, consider the use of robust estimators of the mean and variance.	Normality assumptions do not hold. Generally, robust estimators of the mean and variance can reduce the influence of outliers.
Distribution free	Individual observations are as if obtained by random sampling from populations of almost any kind.	Computations are easy. No external data needed. Populations randomly sampled need not be normal.	Need assumption of independence or symmetry of individual errors arising from random sampling hypothesis.	Can be used if normality assumptions are suspect. Can be used to verify results of parametric tests.	No real disadvantage of these tests. In most cases, power of the test is equivalent or near the parametric counterpart.

These decisions include the effects of interest (model specification — one-way designs, two-way designs, and so forth); whether the classification variables are random, fixed, or nested; whether any interactions (nonadditive effects) are present in the data; how to handle unbalanced designs (unequal sample sizes for the various treatments); and the nature of the error term.

As we can see from this list, ANOVA procedures are not simple but require a great deal of thought. In general, the ANOVA model should follow directly from the sample design used to collect the biocriteria data. The following model illustrates a simple one-way, fixed block design like that described in the upstream/downstream case presented here. The overall model for the ANOVA is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \textbf{(4.9)}$$

where $Y_{i,j}$ = the jth response for the ith site

μ = the population mean

α_i = the effect of site i on Y

$e_{i,j}$ = the error associated with each observation in the data.

The model assumes that the errors are normally distributed with mean 0 and variance σ^2. Based on the model, any observation is composed of an overall mean (μ), a site effect (α), and a random element (e) from a normally distributed population. Hypothesis testing for the ANOVA model is undertaken by calculating the variance associated with model components (sums-of-square differences around the mean effect). A test statistic is formed by comparing the mean square differences associated with a model component to the mean error term. This statistic is distributed as an *F* distribution. Table 4.3 presents an example of this variance breakdown for the simple upstream/downstream model.

Table 4.3.— Analysis of variance results for the case study model.

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	Pr>F
Site	5	146.57	29.31	4.51	0
Error	54	350.67	6.49		
Total	59	497.24			

As seen in the table, the effect of site means is an important indicator of the level of benthic species richness. Therefore, it seems a good idea to explore the relationship among the site means as a method of examining a possible gradient of upstream/downstream differences. Several methods are available for testing the differences between site means. In this example, the method of least significant difference (LSD), Duncan's multiple range test, and Tukey's studentized range test are presented. (A review of these and other multiple comparison methods is in the SAS/STAT Guide for Personal Computers.) Tables 4.4 through 4.6 present the results of these multiple comparison tests.

Table 4.4.—Least significant difference multiple comparison test.

GROUPING		MEAN	*N*	SITE
	A	12.6	10	2
B	A	11.2	10	3
B	A	10.4	10	4
B		9.9	10	1
B	C	8.9	10	6
	C	7.6	10	5

Table 4.5.—Duncan's multiple comparison test.

GROUPING		MEAN	*N*	SITE
	A	12.6	10	2
B	A	11.2	10	3
B	A	10.4	10	4
B	C	9.9	10	1
B	C	8.9	10	6
	C	7.6	10	5

Table 4.6.— Tukey's multiple comparison test.

GROUPING			MEAN	*N*	SITE
	A		12.6	10	2
B	A		11.2	10	3
B	A	C	10.4	10	4
B		C	9.9	10	1
B		C	8.9	10	6
		C	7.6	10	5

In the above tables, sites within a specified grouping are not different at the $\alpha = 0.05$ level of significance.

Nonparametric or Distribution Free Procedures

Distribution free methods for testing multiple sample means are available in much the same format as for parametric tests. The Kruskal-Wallis rank sum test (one-way design) and the Friedman rank sum test (two-way design) are frequently used when the normality assumptions do not hold (see Hollander and Wolfe [1973] for a review of these methods). Multiple comparison methods based on the individual rank scores for each site are available.

Again, the investigator must develop the model to match the experimental design. In the upstream/downstream comparisons of benthic species richness, the Kruskal-Wallis test with a simple one-way model results in a chi-square statistic of 16.38 (Pr < chi-square = 0.006). Again, the upstream/downstream sites appear to differ in the measured biocriteria. Results of the multiple comparison tests using ranks were similar to those presented in the ANOVA model.

A Test for Broad Alternatives

Frequently, investigators are faced with situations in which tests for mean differences or variance differences are not sufficient. For example, investigators may realize that smaller fish are more sensitive to a pollutant than larger fish. In such cases, simple testing for mean differences (in which the mean is calculated without regard to size class) between reference and impacted sites may not suffice. Instead, the measure of toxic effect will be better reflected through changes in the distribution of fish caught at the two sites. Examining the differences in distribution functions among sites may be a more sensitive way to detect effects than relying on population estimates such as the mean and variance.

Statistics designed to detect broad classes of alternatives, as in the scenario presented here, are distribution free tests (i.e., they do not rely on normality assumptions), although they do have parametric counterparts. For a single sample, goodness-of-fit tests to gage the correspondence between an empirical distribution function of observations and a specific probability model or distribution (e.g., normal or lognormal) may be useful. These tests can also be conducted using the chi-square statistic (see Snedecor and Cochran, 1967).

The Kolmogorov-Smirnov Two-Sample Test

Within the biocriteria program investigators will frequently be challenged to evaluate a broad range of differences between two or more populations. The Kolmogorov-Smirnov (KS) two-sample test is easy to implement and can be used to evaluate the relationship between two distribution functions. This test provides graphic and statistical evaluations of two sets of data.

The KS two-sample test involves the development of two cumulative distribution functions (CDFs) to test the hypothesis that each sample was taken from the same population. The test is based on the difference between the empirical distribution functions. The largest difference between the two functions, D_{max}, forms the basis for the test statistic. D_{max} is the maximum vertical distance at any horizontal point between the two CDFs (Fig. 4.1).

To generate a CDF for an individual sample, the data are ordered from lowest to highest, and the rank order of each point determined. Dividing each rank by the sample size results in a cumulative distribution function ranging from 0 to 1 (or 0 to 100 percent, if multiplying by 100). The two samples need not have the same number of observations. Tabled values of the test statistic are available for various sample sizes (Hollander and Wolfe, 1973). The test is both one-sided and two-sided. For the benthic species rich-

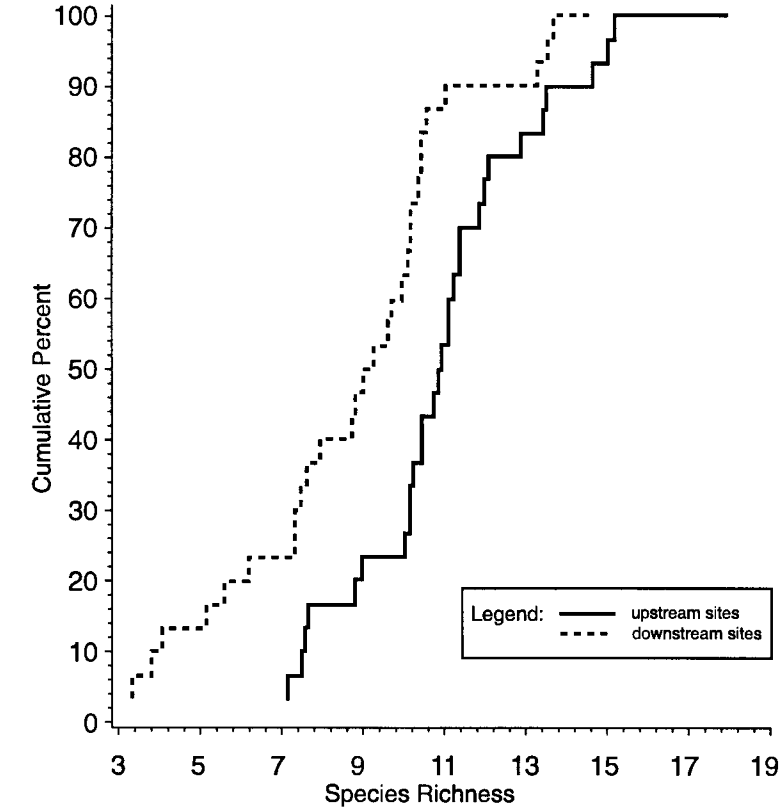

Figure 4.1—Cumulative distribution functions of upstream and downstream sites.

ness example shown here in Figure 4.1, D_{max} is 0.433 (43.3 percent) which occurred at a species richness value of 10.6. The null hypothesis is rejected with a Type I error rate of 0.0072.

Relationship of Survey Design to Analysis Technique

Table 4.7 outlines the relationship between means testing techniques and selected survey designs as described in earlier sections. As a general rule, the data analysis techniques are driven by the survey design. The principle decision points are the number of sites, the available sample size, and the presence or absence of reference sites. However, investigators should not be constrained by the survey design. Data exploration, using any technique that fits the data, is encouraged and can provide insightful results.

Table 4.7.—Survey design and analysis techniques.

SURVEY DESIGN	MEAN DETECTION METHOD
Upstream/downstream: random sampling at single sites using current survey data	*t*-test using an internal value of the variance; Wilcoxon test; with large data sets, a KS two-sample test may be appropriate
Upstream/downstream: random samplings at multiple sites using current survey data	One-way ANOVA using an internal value of the variance; KS two-sample test on merged upstream and downstream data; Kruskal-Wallis rank sum test
Upstream/downstream: random sampling within spatial or temporal strata with one or more sites	Two-way (or more complicated) ANOVA tests; Friedman rank sum test (and other more complicated nonparametric tests)
Impact site data with large off-site external data; for example when determination of impact is not clearly definable or no good upstream reference condition available	External reference distribution tests including the two-sample KS test; *t*-test with external estimate of the variance
Systematic sampling such as random sampling along a transect or nodes of a grid	ANOVA, *t*-tests with internal estimates of the variance, and possibly distribution tests (also note that such designs may be subjected to techniques that demonstrate geographical trends and patterns such as kriging and GIS methods)
Regionally impacted sites with one or more reference sites	Two-sample KS test